CLAIMS

1. A data analysis system for determining a correlation model with a biological condition or a change in the biological condition probabilistically generated with time as an object variable and gene expression levels and/or quantities of intracellular substances as explanatory variables, the system comprising:

an input means for receiving a sample set including data on the biological condition or data from which the biological condition is derived, or data on the change in the biological condition probabilistically generated with time, and the expression levels of a plurality of genes and/or the quantities of intracellular substances;

(A) a selection means for selecting explanatory variables;

(B) a calculation means for calculating a result of cross validation with executing a partial least square method, or for calculating a result of cross validation with executing a partial least square method by using a conversion result as the object variable, the conversion result being obtained by converting a probability that no change occurs by applying a life table according to Kaplan-Meier method or Cutler-Ederer method to the data on the change in the biological condition, based on or without an assumed distribution;

(C) a judging means for judging adoption and non-adoption of the explanatory variables by assessing the result of the calculation by the calculation means; and

(D) a decision means for deciding a partial least square model by continuously improving a function having the result of the cross validation

of the partial least square model as at least one of independent variables by using the selection means (A), the calculation means (B) and the judging means (C).

2. The data analysis system according to Claim 1, wherein the object variable is the biological condition, the data received by said input means are the biological conditions or data from which the biological condition is derived, and the calculation means (B) calculates the cross validation with executing the partial least square method.

3. The data analysis system according to Claim 1, wherein the object variable is the change in the biological condition probabilistically generated with time, the data received by the input means are the data on the change in the biological condition probabilistically generated with time, and the calculation means (B) calculates calculating a result of cross validation with executing a partial least square method by using a conversion result as the object variable, the conversion result being obtained by converting a probability that no change occurs by applying a life table according to Kaplan-Meier method or Cutler-Ederer method to the data on the change in the biological condition, based on or without an assumed distribution;

4. The data analysis system according to Claim 1, 2 or 3, further comprising a final model decision means for constructing a model of a statistical method or multivariate analysis method by using the explanatory variables adopted in the partial least square model determined by the decision means or latent variables of the partial least square model.

5. The data analysis system according to any one of Claims 1 to 4,

wherein the explanatory variables are sequentially adopted or excluded by the selection means.

6. The data analysis system according to any one of Claims 1 to 4, wherein the explanatory variables are selected by using a genetic algorithm by the selection means.

7. The data analysis system according to any one of Claims 1 to 6, wherein the result of the cross validation is calculated with the partial least square method by excluding a sample in the calculation means sequentially.

8. The data analysis system according to any one of Claims 1 to 6, wherein the cross validation is calculated by applying the partial least square method by excluding a plurality of samples in the calculation means sequentially .

9. The data analysis system according to Claim 7 or 8, wherein the calculation means determines a representative value of an error between the object variable showing the biological condition predicted by the gene expression of the excluded sample in each calculation and the object variable showing the biological condition of the excluded sample, and uses the error as an index of the cross validation.

10. The data analysis system according to any one of Claims 1 to 9, wherein the function is the result of the cross validation.

11. The data analysis system according to any one of Claims 1 to 9, wherein the function of the result of the cross validation and the number of the selected explanatory variables is used.

12. The data analysis system according to Claim 5, wherein the decision means repeats discriminative assessment by improving a function

having the result of the cross validation as at least one of the independent variables.

13. The data analysis system according to any one of Claims 1 to 12, wherein the selection means (A) and the calculation means (B) comprise a plurality of computers.

14. A data analysis system further comprising another input means for receiving the correlation model determined in Claims 1, 2, 3 or 4 and the explanatory variables adopted in the model on samples to be predicted, and a means for predicting and discriminating the biological condition of the samples based on the received explanatory variables.

15. The data analysis system according to Claim 2 wherein the object variable is the biological condition represented in nominal scale, ordinal scale or a continuous quantity.

16. The data analysis system according to Claim 4, wherein the statistical method or multivariate analysis method is a regression analysis method applied to a proportional hazard method or parametric distribution.

17. A data analysis method for determining a correlation model with a biological condition or a change in the biological condition probabilistically generated with time as an object variable and gene expression levels and/or quantities of intracellular substances as explanatory variables, the method comprising the steps of:

receiving a sample set including data on the biological condition or data from which the biological condition is derived, or data on the change in the biological condition probabilistically generated with time, and the expression levels of a plurality of genes and/or the quantities of intracellular

substances;

(A) selecting the explanatory variables;

(B) calculating a result of cross validation by executing a partial least square method, or calculating a result of cross validation with executing a partial least square method by using a conversion result as the object variable, the conversion result being obtained by converting a probability that no change occurs when a life table is applied according to Kaplan-Meier method or Cutler-Ederer method to the data on the change in the biological condition, based on or without an assumed distribution;

(C) judging adoption and non-adoption of the explanatory variables by assessing the result of the calculation by the calculation in the calculating step (B); and

(D) deciding a partial least square model by continuously improving a function having the result of the cross validation of the partial least square model as at least one of independent variables by using the selecting step (A), the calculating step (B) and the judging step (C).

18. The data analysis method according to Claim 17, wherein the object variable is the biological condition, the data received in said input step are the biological conditions or data from which the biological condition is derived, and the cross validation is calculated in the calculation step (B) with executing the partial least square method.

19. The data analysis method according to Claim 17, wherein the object variable is the change in the biological condition probabilistically generated with time, the data received in the input step are the data on the change in the biological condition probabilistically generated with time, and

in the calculating step (B) the result of cross validation is calculated with executing a partial least square method by using a conversion result as the object variable, the conversion result being obtained by converting a probability that no change occurs by applying a life table according to Kaplan-Meier method or Cutler-Ederer method to the data on the change in the biological condition, based on or without an assumed distribution.

20. The data analysis method according to Claim 17, 18 or 19 further comprising a step for constructing a model of a statistical method or a multivariate analysis method by using the explanatory variables adopted in the partial least square model determined by the decision means or latent variables of the partial least square model.

21. The data analysis method according to any one of Claims 17 to 20, wherein the explanatory variables are sequentially adopted or excluded in the selection step.

22. The data analysis method according to any one of Claims 17 to 20, wherein the explanatory variables are selected by using a genetic algorithm.

23. The data analysis method according to any one of Claims 17 to 22, wherein the result of the cross validation is calculated with the partial least square method by excluding one sample in the calculation step sequentially.

24. The data analysis method according to any one of Claims 17 to 22, wherein the result of the cross validation is calculated with the partial least square method by excluding a plurality of samples in the calculation step sequentially.

25. The data analysis method according to Claim 23 or 24, wherein in the calculation step a representative value of an error between the object variable showing the biological condition predicted by the gene expression of the excluded sample in each calculation and the object variable showing the biological condition of the excluded sample is determined, and the error is used as an index of the cross validation.

26. The data analysis method according to any one of Claims 17 to 25, wherein the function is the result of the cross validation.

27. The data analysis method according to any one of Claims 17 to 25, wherein the function of the result of the cross validation and the number of the selected explanatory variables is used.

28. The data analysis method according to Claim 21, wherein in the decision step discriminative assessment is repeated by improving the function having the result of the cross validation as at least one of independent variables.

29. The data analysis method according to any one of Claims 17 to 28, wherein the selection step (A) and the calculation step (B) are executed in a plurality of computers.

30. A data analysis method comprising the steps of:

receiving the correlation model determined in Claims 17, 18, 19 or 20 and explanatory variables adopted in the model on samples to be predicted, and

predicting and discriminating the biological condition of the samples based on the received explanatory variables.

31. The data analysis method according to Claim 18 wherein the

object variable is the biological condition represented in nominal scale, ordinal scale or a continuous quantity.

32. The data analysis method according to Claim 20, wherein the statistical method or multivariate analysis method used in the final model decision step is a regression analysis method applied to a proportional hazard method or parametric distribution.

33. A data analysis program executable in a computer for determining a correlation model with a biological condition or a change in the biological condition probabilistically generated with time as an object variable and gene expression levels and/or quantities of intracellular substances as explanatory variables, the program comprising the steps of:

receiving a sample set including data on the biological condition or data from which the biological condition is derived, or data on the change in the biological condition probabilistically generated with time, and the expression levels of a plurality of genes and/or the quantities of intracellular substances;

(A) selecting the explanatory variables;

(B) calculating a result of cross validation by executing a partial least square method, or calculating a result of cross validation with executing a partial least square method by using a conversion result as the object variable, the conversion result being obtained by converting a probability that no change occurs by applying a life table according to Kaplan-Meier method or Cutler-Ederer method to the data on the change in the biological condition, based on or without an assumed distribution;

(C) judging adoption and non-adoption of the explanatory variables

by assessing the result of the calculation by the calculation in the calculating step (B); and

(D) deciding a partial least square model by continuously improving a function having the result of the cross validation of the partial least square model as at least one of independent variables by using the selecting step (A), the calculating step (B) and the judging step (C).

34. The data analysis program according to Claim 33, wherein the object variable is the biological condition, the data received in said input step are the biological conditions or data from which the biological condition is derived, and the cross validation is calculated in the calculation step (B) with executing the partial least square method.

35. The data analysis program according to Claim 33, wherein the object variable is the change in the biological condition probabilistically generated with time, the data received in the input step are the data on the change in the biological condition probabilistically generated with time, and in the calculating step (B) the result of cross validation is calculated with executing a partial least square method by using a conversion result as the object variable, the conversion result being obtained by converting a probability that no change occurs by applying a life table according to Kaplan-Meier method or Cutler-Ederer method to the data on the change in the biological condition, based on or without an assumed distribution.

36. The data analysis program according to Claim 33, 34 or 35 further comprising a step for constructing a model of a statistical method or multivariate analysis method by using the explanatory variables adopted in the partial least square model determined by the decision means or latent

variables of the partial least square model.

37. The data analysis program according to any one of Claims 33 to 36, wherein the explanatory variables are sequentially adopted or excluded in the selection step.

38. The data analysis program according to any one of Claims 33 to 36, wherein the explanatory variables are selected by using a genetic algorithm.

39. The data analysis program according to any one of Claims 33 to 38, wherein the result of the cross validation is calculated with the partial least square method by excluding a sample in the calculation step sequentially.

40. The data analysis program according to any one of Claims 33 to 38, wherein the cross validation is calculated by applying the partial least square method by excluding a plurality of samples in the calculation method sequentially.

41. The data analysis program according to Claim 39 or 40, wherein in the calculation step a representative value of an error between the object variable showing the biological condition predicted by the gene expression of the excluded sample in each calculation and the object variable showing the biological condition of the excluded sample is determined, and the error is used as an index of the cross validation.

42. The data analysis method according to any one of Claims 33 to 41, wherein the function is the result of the cross validation.

43. The data analysis method according to any one of Claims 33 to 41, wherein the function of the result of the cross validation and the number

of the selected explanatory variables is used.

44. The data analysis program according to Claim 37, wherein in the decision step discriminative assessment is repeated by improving the function having the result of the cross validation as at least one of independent variables.

45. The data analysis program according to any one of Claims 33 to 44, wherein the selection step (B) and the calculation step (C) are executed in a plurality of computers.

46. A data analysis program comprising the steps of:

receiving the correlation model determined in Claims 33, 34, 35 or 36 and the explanatory variables adopted in the model on samples to be predicted, and

predicting and discriminating the biological condition of the samples based on the received explanatory variables.

47. The data analysis method according to Claim 34 wherein the object variable is the biological condition represented in nominal scale, ordinal scale or a continuous quantity.

48. The data analysis method according to Claim 36, wherein the statistical method or multivariate analysis method used in the final model decision step is a regression analysis method applied to a proportional hazard method or parametric distribution.

49. The program according to Claim 37, wherein the explanatory variables are not included at all in an initial state in the selection step.

50. The program according to Claim 37, wherein the full explanatory variables are included in the initial state in the selection step.

51. The program according to any one of Claims 37 to 50, wherein the biological condition comprise a measured value representing a type of disease, a measured value representing degree of critical of disease, a result of diagnosis representing a type of disease, a result of diagnosis representing degree of critical of disease, or a value derived from one of them.

52. A computer readable recording medium which records the program according to any one of Claims 33 to 48.

53. An intracellular substance measuring device, wherein the device detects expression of at least one gene selected substantially in a gene group of gene bank accession numbers U15085, M23452, X52479, U70426, H57330 and S69790, a detection method thereof, or a method for examining degree of critical of diffuse large cell lymphoma based on the detection.

54. The intracellular substance measuring device, the measuring method thereof, or the examining method according to Claim 53, wherein expression of at least one gene selected substantially from a gene group of gene bank accession numbers U03398, M65066, AK001546, BC003536, X00437, U12979, H96306, AA830781 and AA804793 is further detected.

55. An intracellular substance measuring device, wherein intracellular substances containing gene products comprising substantially genes of gene bank accession numbers AA598572, AA703058 and AA453345 are detected, a measuring method thereof, or a method for examining degree of critical of breast cancer based on the detection.

56. The intracellular substance measuring device, the measuring method thereof, or the examining method according to Claim 55, wherein

intracellular substances containing substantially at least one gene product selected from a gene group of gene bank accession numbers AA406242, H73335, W84753, N71160, AA054669, N32820 and R05667 are further detected.

57. An intracellular substance measuring device wherein intracellular substances containing gene products comprising substantially genes of gene bank accession numbers W84753, H08581, AA045730 and AI250654 are detected, a measuring method thereof, and a method for examining recurrence of breast cancer based on the detection.

58. The intracellular substance measuring device, the measuring method thereof, or the examining method according to Claim 57, wherein the intracellular substances containing substantially at least one gene product of genes selected from a gene group consisting of gene bank accession numbers AA448641, R78516, R05934, AA629838 and H53037 are further detected.

59. An intracellular substance measuring device, wherein intracellular substances containing gene products substantially comprising genes of gene bank accession numbers AA434397, T83209, N53427, N29639, AA485739, AA425861, H84871, T64312, T59518 and AA037488 are detected, a measuring method thereof, and a method for examining recurrence of breast cancer based on the detection.

60. The intracellular substance measuring device, the measuring method thereof, or the examining method according to Claim 60, wherein intracellular substances containing gene products substantially of a gene of gene bank accession number AA406231 are detected further.

61. An intracellular substance measuring device, wherein intracellular substances containing gene products substantially of genes of gene bank accession numbers H11482, T64312 and AA045340, a measuring method thereof, and a method for examining recurrence of breast cancer based on the detection.